Density Estimation is a problem of reconstructing the pdf using a set of given data points. Suppose we observe $x_1, \ldots, x_n$ and we want to recover the underlying PDF generating our data set. We assume the data comes from a continuous prob. distribution.

In parametric, we assumed the distribution to be some known distribution and estimate its parameters but here we are not assuming any distribution, rather inventing it.

* Estimating CDF :

Assume that X follows a continuous distribution with CDF F. Then, for $h > 0$ $\mathbb{P}(x - \frac{h}{2} < X < x + \frac{h}{2}) = \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y)\,dy \approx h \cdot f(x)$ (as $x + \frac{h}{2} - x + \frac{h}{2}$) when h is small.

Hence, $\hat{f}(x) = \dfrac{F(x + h/2) - F(x - h/2)}{h}$ (1).

* Empirical CDF :

Let $x_1, x_2, \ldots, x_n$ be iid random variables then the empirical CDF is given by : $F_n = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i \leq x)$ (2).

$\hat{f}(x) = \dfrac{1}{nh} \sum_{i=1}^{n} \mathbb{1}(x - \frac{h}{2} \leq x_i \leq x + \frac{h}{2})$ (EDE) ;

1) $\hat{F_n}$ is UE of true distribution
2) $\mathrm{Var}(\hat{F_n}) = \dfrac{1}{n} F(x)(1 - F(x))$
3) $\hat{F} \xrightarrow{\mathbb{P}} F(x)$

3) $F_n \xrightarrow{n} F(n)$

$$\mathbb{E}(\hat{F_n}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\mathbb{1}(X_i \le n))$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} \mathbb{1}(X_i \le n)\, f(y)\, dy$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{n} f(y)\, dy$$

$$= \frac{n F(y)}{n}$$

$$= F(y)$$

Drawback : A continuous distribution is being estimated by a discrete function.

The estimate requires a large data sample.

* Def : Kernel Density Estimation (KDE)

A Kernel is called a kernel function if $K(n) \ge 0$, $\int_{-\infty}^{+\infty} f(n)\, dn = 1$ & $K(-n) = K(n)$ e.g gaussian kernel.

$$\hat{f}(n) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{n - x_i}{h}\right), \quad h : \text{bandwidth} \quad x_1, \dots, x_n : \text{Sample of } n \text{ observations}$$

However, this is a biased estimator and the variance is complicated.

* NW Estimator & KDE (Nadaraya & Watson) :

They invented Kernel regression : $\hat{f}(n) = \sum y_i w_i$

where $w_i = \frac{K\left(\frac{n-n_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{n-n_i}{h}\right)}$

* Kernel Density Estimation & Probalistic Neural Network

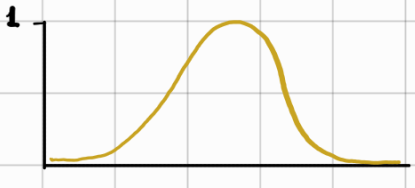
- Y Numerical : LR, LASSO, Ridge, Spline, Polynomial, Elastic Net (Regression Models).
- Y Categorical : Logistic, Bradly Terry, PNN (Classification models).
- Y Count : Poisson (Count Regression)
- Y Time Series : AR, ARIMA, BSTS, Bass (Forecasting Models)
- Y Time to event : Cox PH (survival Modelling)

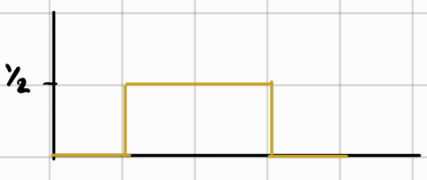
* Examples :

1) Gaussian Kernel : $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$

2) Epanechnikov Kernel : $K(u) = \frac{3}{4} \max\{1-u^2, 0\}$ (Quadratic)

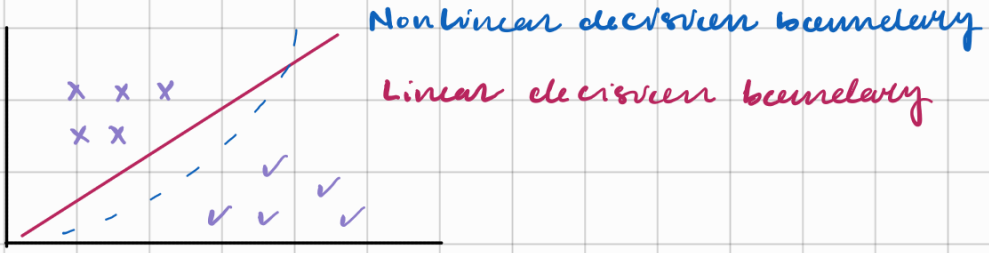
3) Uniform Kernel : $K(u) = \begin{cases} 1/2 & , |u|<1 \\ 0 & , \text{o.w} \end{cases}$





* PNN : 1) Useful for binary / multi-class classification problem

2) Handles nonlinear decision boundaries



3) Provides probabilistic outputs for classification

* Structure of a probabilistic Neural Network :

1) Input Layer : Consists of features of the data

2) Pattern Layer : Each neuron in this Layer contains one neuron for each training sample. This layer computes a similarity between test input and training samples using kernel.

3) Summation Layer : Aggregates the output from pattern layer neurons for each class summing the similarity (kernel values)

4) Decision Layer : Calculates prob of test input belonging to each class

Feature 1 (x_1)	Feature 2 (x_2)	Output	
2.0	2.0	O	
2.2	2.2	O	Imbalanced
2.4	2.4	O	data 80% - 20%
2.6	2.6	O	
2.8	2.8	O	Imbalance Ratio:
1.8	1.8	O	$IR = \frac{8}{2} = 4$
1.9	2.1	O	$IR = \frac{\#\text{ Majority}}{\#\text{ Minority}}$
2.3	2.1	O	

4.0	4.0	1
4.1	4.1	1

Test point : $(x_1, x_2) = (3.5, 3.5)$

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right)$$

$\sigma = 1$

X_1	X_2	Output	$K(x, x_i)$	Summing	Decision
2.0	2.0	0	0.105	1.98	1.98
2.2	2.2	0	$\exp\left(-\frac{2(3.5-2)^2}{2}\right)$		Class 0
2.4	2.4	0	+		
2.6	2.6	0	+		
2.8	2.8	0	+		
1.8	1.8	0	+		
1.9	2.1	0	+		
2.3	2.1	0			
4.0	4.0	1	+	1.47	
4.1	4.1	1			

Probability : $\mathbb{P}(x \in \{0\}) = \frac{1.98}{1.47 + 1.98} = 0.57$

Really Close!

$\mathbb{P}(x \in \{1\}) = \frac{1.47}{1.47 + 1.98} = 0.43$

Fit a Skew Normal to the train

Skew normal : $Sk(x, x_i) = K(x, x_i)\, \Phi\left(\frac{\hat{\alpha}\|x - x_i\|}{\sigma}\right) \quad \alpha \in [-6, 6]$

Skew PNN : 0 : 0.0347 $\quad \mathbb{P}(x \in \{0\}) = 0.16$

1 : 0.185 $\quad \mathbb{P}(x \in \{1\}) = 0.84$

$\Rightarrow n \in \{1\}$